View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Yonsei University Medical Library Open Access Repository

보건정보통계학회지 제40권 제1호
ISSN 2287-3708(Print) ISSN 2287-3716(Online)
Journal of Health Informatics and Statistics (JHIS)
2015; 40(1): 75-86

Propensity score model 구축에서 상관성을 고려한 변수선택

박성훈[1], 송기준[1]†

[1]연세대학교 의과대학 의학통계학과

Variable Selection for Propensity Score Models Considering the Correlations between Covariates

Seong Hun Park[1], Kijun Song[1]†

[1]Department of Biostatistics, College of Medicine, Yonsei University

Abstract

Objectives: In the covariate selection for propensity score model (PSM), including all the covariates that can be observed has been recommended. However, there are problems that appear multi collinearity and do not obtain the matching number needed using over fitted propensity score model. In this study, we studied the method of variable selection for PSM considering the correlations between covariates.

Methods: All the covariates were classified according to the relation with treatment and outcome and generated considering the correlations each other. We examined the odds ratio and MSE (mean squared error) of PSM and the matching number of simulated data.

Results: When there are correlations among covariates included in PSM, the matching number decreased as the correlation of covariates was stronger. Also, the larger the strength of correlation among covariates was, the smaller MSE was and the matching number was.

Conclusions: When including covariates in PSM, we found that it is more efficient to examine the correlation of covariates, treatment variable, and outcome variable than using all the covariates observed.

Keywords: Propensity score, Matching, Simulation, Variable selection

[Submitted: 2015년 02월 02일, Revised: 2015년 02월 21일, Accepted: 02월 23일]

† Corresponding Author: Kijun Song, PhD
Department of Biostatistics, College of Medicine, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea. Tel:+82-2-2228-2491
E-mail: biostat@yuhs.ac

1. 서 론

보건의학분야에서 사용되어지는 관찰연구 중에서 두 군을 비교하는 연구는 대부분 비무작위 시험을 기준으로 되어 왔다. 이러한 비무작위 시험은 선택편의(selection bias)를 통제할 수 없는 문제점을 가지고 있다. 선택편의는 처리집단과 대조집단 간의 이질성으로 인해 발생되는 것으로, 처리와 결과의 인과관계에 대한 올바르지 못한 추론을 하거나 처리효과를 과소 혹은 과대 추정하는 오류를 발생시키게 된다. 따라서 처리군에 대응되는 대조군을 선정할 때 공변량(covariate)들의 불균형(unbalanced)을 통제할 수 있는 대상이 선정 되어야 할 필요성이 있다. 두 집단에서 구조적인 군 간의 차이가 발생하지 않도록 하기 위해 처리변수와 결과변수가 이분형 자료인 경우, 로지스틱회귀모형을 이용하여 균등하게 만들어야 할 공변량들을 모형에 포함시킨 PSM (propensity score model)을 구축하여 두 군간 개체들을 대응시키는 방법이 많이 사용되어지고 있다 [1]. 이러한 PSM을 이용할 때 기존의 많은 연구들에서는 측정할 수 있는 가능한 많은 공변량들을 PSM에 포함시켜 잠재적 혼란변수를 통제하는 것을 제안해왔다 [2]. 그런데, 최근의 연구들에서 가능한 많은 공변량들을 PSM에 포함시키는 것이 불균형을 줄이는데 반드시 효율적이지는 않다는 결과를 제시하였다 [3,4]. Alan et al. [3]의 연구에서는 결과변수에만 상관성을 가지는 공변량은 PSM에 포함이 되어도 편차를 증가시키지 않으며, 처리변수에만 상관성이 있는 공변량은 처리효과의 분산을 증가시킨다고 하였다. Austin et al. [4]의 연구에서는 처리-결과변수와 관계있는 공변량만을 포함하여 PSM을 이용할 것을 제안하였다. 본 연구에서는 이러한 최근의 연구들을 바탕으로 PSM에 포함시킬 공변량의 선택 기준을 제시하고자 한다. Austin et al. [4]의 연구실험설정을 기본으로 이를 확장하여 각 범주에서 2개의 공변량을 생성하여 모의실험을 수행하고, 나누어진 각 범주별로 상관성(correlation)의 조합을 고려하여 PSM에 포함할 공변량의 특성과 상관성을 제시하고자 한다. Austin et al. [4]의 연구에 적용된 모의실험 설정틀은 다음과 같이 정리할 수 있다. 먼저, 데이터를 생성하고 몬테카를로(Monte Carlo) 방법을 적용하여 통계량 평균값을 구한다. 데이터는 특정 분포에서 임의적으로 공변량들을 생성하고, 생성된 공변량들을 이용하여 처리변수, 결과변수간의 연관성과 공변량 간의 상관성 여부에 고려하여 처리변수와 결과변수를 베르누이시행을 통해 생성한다. PSM은 로지스틱회귀를 이용하며, 포함되는 변수의 특성에 따라 각기 다른 PSM을 구축한다. 각 모형으로 PSmatching을 수행한 표본에서 오즈비(odds ratio), MSE (mean squared error), matched number, 표준화차이 계수(standardized differences) 값을 구하고 구해진 통계량들의 평균을 구한 후 이 값들을 통하여 각 모형에 포함된 변수들의 상관성을 비교한다.

2. 연구 방법

1) Propensity score

Propensity score는 관찰된 공변량들이 주어졌을 때, 특정 시험군에 할당되도록 영향을 주는 독립변수에 대한 조건부 확률로써 다음과 같이 정의된다.

$$e(x) = pr(z = 1|x)$$

여기에서 z는 처리수준을 나타내며 x는 측정되어진 공변량이다.

강한 무관성의 가정이 성립할 경우, propensity score를 이용하여 두 군의 개체들을 matching을 하여 각 군의 측정 가능한 특성들은 동일한 분포를 가지게 된다. 이러한 것은 무작위 실험과